Cite as: Malthouse, E., Russell, S., Liang, Y., & Hills, T. (2021). The influence of exposure to randomness on lateral thinking in divergent, convergent, and creative search. *Cognition* (in press)

Title: <u>**The influence of exposure to randomness on lateral thinking in divergent, convergent, and creative search**</u>

Authors: Eugene Malthouse, Serena Russell, Yuanjing Liang, Thomas Hills

Affiliation: University of Warwick, Department of Psychology

Abstract

Exposure to random stimuli has often been suggested to help unlock problem-solving abilities and creativity, helping us to see problems differently and imagine new possibilities. Equally, randomness is widely used in computer science to escape local maxima and find effective solutions to intractable problems. However, randomness has rarely been used as a formal aid in human decision making or investigated in controlled experimental settings. In this pre-registered study, we tested the effect of extraneous random stimuli using Wikipedia's random page generator on 592 British participants' performance across three online tasks: one 'convergent' forecasting task and two 'divergent' fluency tasks. We found no improvement associated with the treatment and often significant impairment. A Bayesian meta-analysis of the tasks finds strong support for the null hypothesis. We conclude that stimulating lateral thinking through random stimuli is non-trivial and may require such stimuli to be sufficiently task-related or 'optimally random'.

Key words: creativity, randomness, decision making, judgment

Introduction

Popular strategies for problem solving often invoke the idea of out-of-the-box thinking (Weisberg & Markman, 2009; de Bono, 2010; Sloane, 2017). The claim is that creatively solving hard problems typically requires insight (Ansburg & Dominowski, 2000) – a realisation of a different way of thinking about the problem (Wertheimer, 1959). Of course, if one knew how to think about the problem differently, then the problem would not be hard. So how does one suddenly recognise the importance of unused information (e.g., that a coin stamped 521 BC could not have been printed at that time), overcome functional fixedness (e.g., realising that the box in the Candle-Box problem can be used as a stand for attaching the candle to the wall), and realise which constraints to relax (such that one can solve the Nine-Dot problem with four lines by literally drawing the line outside the box)?

Stories of the scientific discoveries of buoyancy by Archimedes, gravity by Newton, and penicillin by Fleming, each rely on some form of serendipitous random accident (Christian & Griffiths, 2016). One potential solution therefore might be the deliberate use of external random stimuli, an approach that has previously shown promise in stimulating human creativity in certain experimental settings (e.g. Finke et al, 1992; Proctor, 1993). This research suggests that such stimuli can disrupt preconceived patterns of thought (Beaney, 2005) or trigger novel associations (Simonton, 2011). In other words, internal randomness may be adaptively upregulated – e.g., by reducing low level inhibition – to facilitate the generation of alternatives (Carson, Peterson, & Higgins, 2003; Chrysikou, 2019; Hills, 2019; Sweller, 2009). Could exposure to an external source of randomness therefore achieve a similar aim in both convergent and divergent thinking tasks? In other words, could the absolute randomness of external stimuli increase the relative randomness of internal cognitive processes?

Theories of creativity frequently invoke the notion of internal exploration (Hommel, 2012). One of the first, proposed by Campbell (1960), suggested that creativity arises from the human mind's selection of the best ideas that it randomly generates. It follows that introducing randomness via external stimuli might further stimulate the mind's idea-generation capability, increasing the pool of ideas available for selection. Relatedly, Vul and Pashler (2008) proposed that people make decisions and generate solutions by sampling probabilistically from their memory. External randomness in this case may help to broaden the sampling regime.

Computer science often invokes a similar approach in which randomness is used in simulations to avoid local maxima and find better solutions to intractable problems. A useful example is simulated annealing, in which 'high temperature' perturbations with slow cooling lead asymptotically to better identification of the global maxima (Kirkpatrick et al, 1983). According to Christian and Griffiths (2016), perturbing people out of local maxima might similarly help them to identify global maxima. They cite Brian Eno's 'Oblique Strategies' and Wikipedia's 'random article' function as possible methods for achieving this creativity-boosting perturbation. De Bono's book, *Lateral Thinking* (2010), devotes a chapter to 'random stimulation' with the following instruction:

> "...one can deliberately generate external stimulation which then acts on the idea from the outside. This is how random stimulation works...With random stimulation one uses any information whatsoever. No matter how unrelated it may be no information is rejected as useless. The more irrelevant the information the more useful it may be" (p.169).

De Bono (2010) argues that such random stimulation can prompt people to realise new patterns and restructure problems in new ways, and thus to engage in the process of "lateral thinking". However, the effectiveness of such methods across different types of tasks remains unclear.

Researchers since Guilford (1967) have identified two main cognitive processes that we bring to bear on different tasks requiring creative input: divergent and convergent thinking. Divergent thinking refers to the process of flexibly generating numerous ideas for a problem that has multiple possible solutions (Runco, 2010). We might think divergently when designing a new product, for example, or writing a fictional story. In contrast, we use convergent thinking for well-defined problems with unique correct solutions, such as spelling an obscure word or calculating company profits from the previous quarter. Convergent thinking has therefore been described as more focused, because it involves distilling all available and relevant data into a single solution (Guilford, 1950). Random stimulation may potentially be useful for both of these kinds of cognitive processes.

In the present study, the solutions requested from participants were forecasts relating to COVID-19, a task which similarly requires envisioning multiple different outcomes and evaluating these critically before converging on a single best estimate (Byrne, Shipman, & Mumford, 2010). While divergent and convergent thinking processes are often described as distinct phenomena (de Vries & Lubart, 2019), creative production typically requires sequential use of both types of thinking: for example, an initial divergent exploration of ideas followed by a convergent identification of the most suitable solution (Cropley, 2006; Zhang et al, 2020). Might exposure to random stimuli therefore support both divergent and convergent thinking processes, and therefore enhance creativity and judgment?

To our knowledge, there has been limited experimental investigation into the use of exposure to randomness as a formal aid in divergent or convergent tasks. In this pre-registered study (https://osf.io/ns6v2) we tested whether an exogenous randomness tool could help participants to improve their performance in both types of tasks. The tool itself consisted of randomly selected Wikipedia articles with which participants were asked to engage. This treatment was used in all three studies described below. It was designed to stimulate out-of-the-box lateral thinking by derailing individuals' habitual trains of thought. In other words, we imagined that in their cognitive search for answers, the random stimuli might prompt participants to think laterally – or to think about more different possible solutions – rather than searching further within the same constrained 'cognitive hyperspace' (Acar & Runco, 2015). While we did not explicitly instruct these participants to integrate the presented random elements into their responses, we did suggest that they may serve as a decision aid in the task. To ensure a consistent experience across treatment groups, we used similar suggestive language about the task given to participants in the control condition.

Study 1 was a convergent task in which subjects made quantitative forecasts about the impact of COVID-19. The literature on forecasting has burgeoned since the publication of *Superforecasting* (Tetlock & Gardner, 2016) (e.g., Hubbard, 2020; Katsagounos et al, 2020), but the potential benefit of introducing randomness into the forecasting process remains unclear. Further, we wanted to investigate how the usefulness of our randomness tool would compare to a more traditional accuracy-boosting strategy called 'consider the opposite' (Lord et al, 1984). This involves prompting people, after they have made a first estimate, to consider whether any of their underlying assumptions might have been wrong before making a revised second estimate – a process that has been shown to improve estimate accuracy (Herzog & Hertwig, 2009). We hypothesised that participants who used our randomness tool would generate more accurate forecasts than both those in the control group and those 'considering the opposite'. In Study 2 and 3, we presented participants with divergent tasks involving the search for ideas relating to COVID-19 and public policies, respectively. Here, we hypothesised that participants using the same randomness tool would generate a higher number of more creative ideas than those in the control group. In sum, the present studies investigated the value of using a randomness tool as a formal aid for enhancing human judgment and creativity.

For all three studies, participants were sourced from Prolific Academic, an online platform designed for experimental research recruitment (Palan & Schitter, 2018). All participants were UK residents over the age of 18 and received fair payment that equated to approximately £7.50 per hour. This rate, and the experiment as a whole, was approved by the University of Warwick's Psychology Department Research Ethics Committee. We conducted power calculations based on predicted effect sizes and used sample sizes from similar studies (e.g. Herzog & Hertwig, 2009; Herzog & Hills, 2019) as guidance for determining our minimum sample sizes for our three studies. For Study 1 and 2, our sample sizes were chosen to exceed that in Herzog and Hertwig

(2009). For Study 3, which took three times as long to complete, we were pragmatically constrained to collect one-third of the sample size. The Bayesian meta-analysis across the three studies provides a definitive statement regarding the adequacy of our sample sizes. Each study followed a between-subjects design, with participants randomly assigned to treatment groups. In all three studies, participants responded over two rounds either side of a treatment or control exercise, allowing us to isolate and evaluate the effect of the randomness tool.

Study 1: Lateral thinking and convergent search

Methods

214 participants (67.3% female) were asked to produce a total of six forecasts related to COVID-19 (see Supplemental Materials). In short, these were: 1a) the cumulative total number of COVID-related global deaths that would be reported two weeks after the survey completion date; 1b) the cumulative total one year after the survey; 2a) the cumulative total number of countries suffering more than 5,000 COVID-related deaths in two weeks' time; 2b) the cumulative total in one year's time; 3a) the UK unemployment rate between March and May 2020; and 3b) the UK unemployment rate between March and May 2021. For each forecast, participants were provided with the most up-to-date figures available for COVID-19 from the European Centre for Disease Prevention and Control (https://www.ecdc.europa.eu/en) and for UK unemployment from the Office for National Statistics (https://www.ons.gov.uk/). Subjects were asked to produce a best estimate for each forecast, as well as lower and upper bound estimates such that they were 90% confident that the true value would lie between these figures.

Participants were randomly assigned to one of three treatment conditions. For each forecast, they first provided a best estimate bracketed by upper and lower bounds such that they were 90% confident that the true value would lie between these figures. They then received their treatment, after which they were asked to revise their initial estimates (without being able to see these previous estimates on screen).

In our lateral thinking treatment, we showed participants screenshots from four random Wikipedia pages and asked them to summarise each page in 10 words or less. We ensured that different participants saw different combinations of pages by drawing a random sample of four pages from a pool of 100. Pages were collected indiscriminately in advance using Wikipedia's random page function (http://en.wikipedia.org/wiki/Special:Random), with subjects ranging from species of moth to professional cyclists. The consider-the-opposite treatment mirrored the format previously used by Herzog and Hertwig (2009). Participants were asked to assume that their first estimates were off the mark; to think about a few reasons why that could be; to identify which assumptions could have been wrong; and to consider whether their estimates were consequently too low or too high. In our control treatment, subjects completed four mental arithmetic questions, examples of which can also be found in the Supplemental Materials, a task format we borrowed from Niederle and Vesterlund (2007).

For each forecast, subjects' responses were excluded only on the basis of mathematical impossibility: e.g., if their lower bound estimate was greater than their best or upper bound estimate. In addition, for the first two forecasts we excluded responses that fell below the up-to-date figure shown to participants (i.e. 649,358 and 22, respectively). These exclusion criteria left 154 valid responses for the first forecast (1a and 1b above), 170 for the second (2a and 2b above), and 195 for the third (3a and 3b above) – a difference explained mostly by errors associated with typing larger numbers in the first forecast and participants' greater familiarity with the task by the time they reached the third forecast. Before evaluating the treatment effects, we first conducted analyses of variance (ANOVAs) on participants' first round (pre-treatment) responses to check for any differences between the groups. No significant differences were found in this or any of the other studies (see Supplemental Table 1).

Results

Our analysis of Study 1 responses focused on forecast accuracy between treatment conditions, which we assessed via two separate methods. First, for each forecast we calculated the mean absolute percentage error (MAPE) of participants' first (pre-treatment) estimates, revised (post-treatment) estimates, and the average of these two estimates. Put simply, the MAPE is the difference between the forecast and the true value (the error) as a percentage of the true value. We chose this method because it is scale-independent, enabling us to compare an individual's accuracy across all three forecasts, and because it is a widely used method of assessing forecast accuracy (Hanke et al, 2001).

We ran an analysis of covariance (ANCOVAs) to evaluate the effect of the treatment on the average accuracy of participants' revised estimates, controlling for the accuracy of individuals' pre-treatment estimates. We did not detect an overall significant main effect (p = .24) of treatment (results from individual forecasts can be found in Supplemental Table 2). We therefore concluded that neither the random nor the consider-the-opposite treatments helped people to improve their forecast accuracy any more than did the control treatment. Figure 1 illustrates this similarity in the distributions of changes in forecast error from pre-treatment to post-treatment estimates between groups.

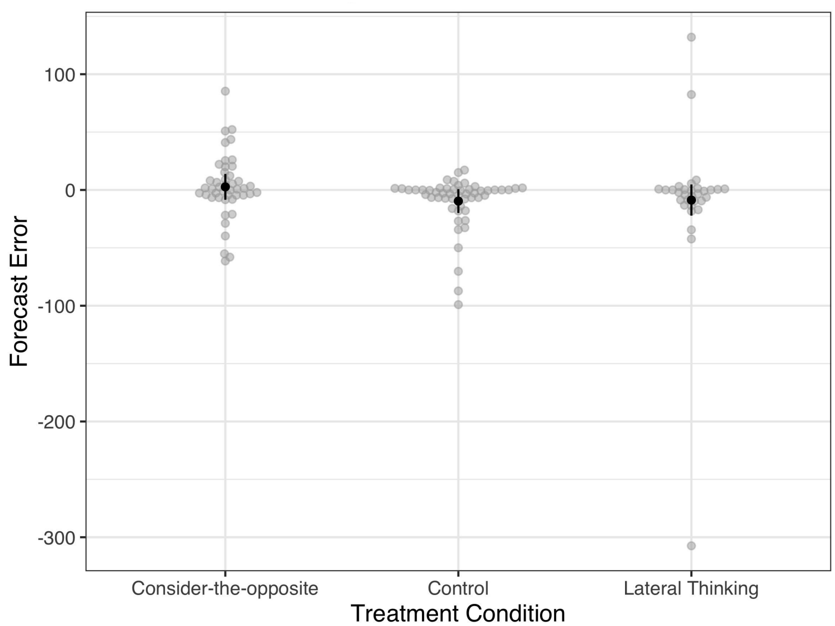


Fig. 1. Forecast Error Change

Average forecast error change before and after treatments. Forecast error is represented by the mean absolute percentage error (MAPE), with negative values signalling a reduction in error (and therefore improved accuracy). Points in the background represent the raw data, with black points showing mean responses and black error bars representing 95% between-subjects confidence intervals.

Following this, we investigated whether the average of an individual's two estimates was more accurate than their first estimate alone. Comparing these figures showed that taking this average of two estimates consistently resulted in higher accuracy for the majority of participants. For the first forecast (on the number of future deaths related to COVID-19) 59.7% of participants improved their accuracy, with an average improvement across all participants of 0.88%. For the second (on the number of countries experiencing more than 5,000 deaths related to COVID-19) 66.5% of participants improved their accuracy by 0.18%, on average. And for the third (on UK unemployment) 74.8% did, with an average improvement of 8.67%. However, there was no significant difference between treatment groups ($F(2, 126) = 1.33$, $p = .27$). These results indicate that while the average of an individual's two forecast estimates tended to be more accurate than their first estimate, there was no marginal benefit to using our lateral thinking Wikipedia tool or employing a consider-the-opposite strategy during this process.

The second approach to evaluating participants' accuracy was to look at whether their 90% confidence bounds bracketed the true value in each forecast. Overall, participants' pre-treatment intervals were poorly calibrated, as shown in Supplemental Figure 1. For the first forecast, overall just 41.1% of valid pre-treatment bounds successfully bracketed the true value. For the second and third forecasts, the equivalent figures were 78.9% and 15.3% respectively. This follows previous evidence that individuals are often unduly overconfident about the precision of their knowledge (e.g. Russo & Schoemaker, 1992). This overconfidence was not attenuated by any of our treatments: across all groups, post-treatment bracketing success was even lower for the first two forecasts (38.3% and 72.4%, respectively), and only slightly higher (20%) for the third forecast. In short, participants' confidence bounds remained too narrow, regardless of treatment condition.

Study 2: Lateral thinking and divergent search I

Methods

263 participants (64.6% female) were divided randomly into two treatment groups and were first given three minutes to generate as many predictions as possible about how society might change after the COVID-19 pandemic. Because we wanted to explore the breadth (rather than the depth) of their ideas, participants were instructed to limit each response to 10 words or less. They then received either the lateral thinking or the control treatment (identical to those used in Study 1), after which they were given a further three minutes to generate as many additional new predictions as possible.

Responses were scored independently, blindly, and quantitatively by both human judges and SemDis (http://semdis.wlu.psu.edu/) – a computational linguistic model that calculates semantic distances between a divergent task subject and a participant's responses (Beaty & Johnson, 2021). These methods enabled us to calculate fluency and creativity scores for each participant in each round (i.e., before and after the treatment). The fluency score was calculated by human judges as the number of valid and distinct predictions generated by a participant, a prominent method used in the creativity research literature (Plucker & Makel, 2010). To count as valid, responses had to clearly indicate a prediction with a measurable outcome: 'higher obesity rates' would be valid, for example, but 'health' would not. Creativity, on the other hand, was scored by both human judges and SemDis. Following Wilson et al (1953), human judges rated each valid response on a 5-point scale according to its uniqueness (1 = not at all unique, 5 = highly unique) according to two criteria: originality (how frequently it appeared in the sample) and remoteness (how conceptually distant it was from common responses). While some more recent research supports the use of this subjective method (e.g., Dumas et al, 2020) the robustness of judge-based scores has also been questioned (e.g., Maio et al, 2020). We therefore also used SemDis to compute a more objective creativity score for every participant in each round based on the total semantic distance between their responses and the task subject ('pandemic'). For maximum validity, we used the platform's multiplicative (rather than additive) compositional model and the mean semantic distance score from the five different semantic spaces available on the SemDis platform for each participant response (Beaty & Johnson, 2021). This score served as a more objective and reliable complement to the ratings given by human judges (Beaty & Johnson, 2021; Dumas et al, 2020). In sum, the above scoring methods enabled us to evaluate and compare fluency and creativity scores between rounds and treatment conditions.

Results

Our analysis of Study 2 data explored whether the lateral thinking treatment (compared to a control treatment) enhanced people's fluency and creativity. Responses were rated for both fluency and subjective creativity by human judges, and for objective creativity by SemDis. Responses were rated for fluency and creativity by human judges and SemDis and were then aggregated to create pre- and post-treatment fluency and creativity scores for each treatment group, enabling us to compare performance between rounds (see Figure 2).

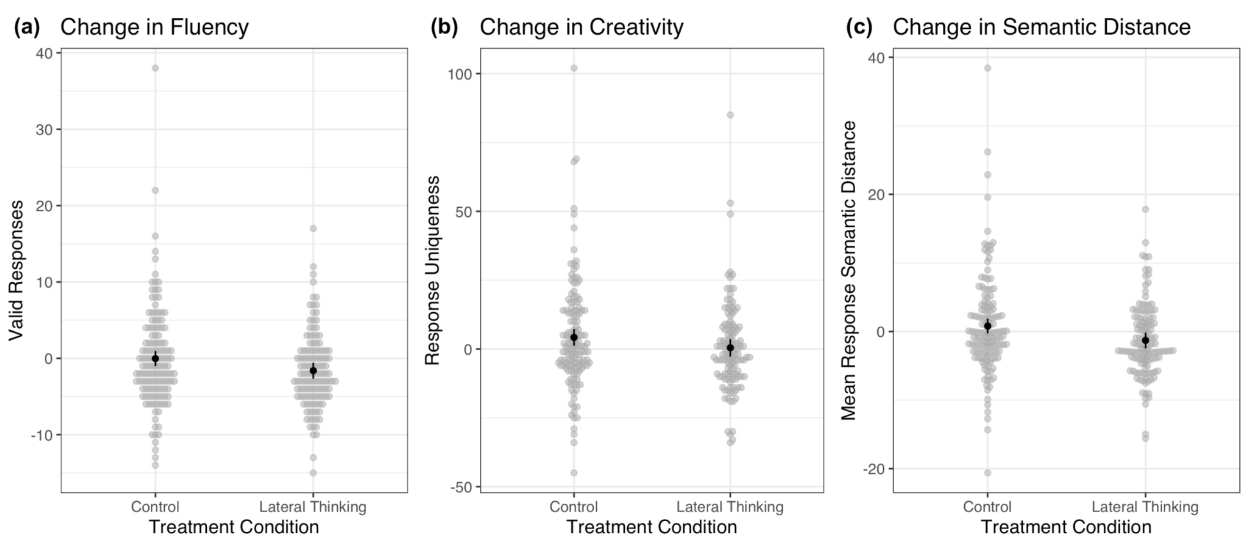


Fig. 2. Change in Fluency and Creativity Scores, Study 2

Change in fluency (a) and creativity scored by human judges (b) and the semantic distance model (c) between rounds. The plots illustrate the change in participants' response fluency and creativity after (relative to before) the treatment. Grey points show the raw data, black points show the mean scores and error bars show 95% between-subjects confidence intervals.

To test for a treatment effect on fluency scores, we ran an ANCOVA (controlling for pre-treatment scores) which indicated a statistically significant effect (p = .020). Contrary to our hypothesis, however, follow-up tests showed that fluency was on average *lower* among participants in the lateral thinking condition than those in the control condition ($t(264) = 2.34$, p = .020).

We ran equivalent ANCOVAs (controlling for pre-treatment scores) to test for a treatment effect on the creativity scores generated by our human judges and SemDis, which were strongly correlated ($r(261) = .72$, $p < .001$). The treatment effect on human-scored creativity was not significant (p = .089), but for creativity in terms of semantic distance it was (p = .009). A follow-up test similarly indicated that the control group's post-treatment SemDis scores were on average more creative than those in the lateral thinking condition ($t(260) = 2.65$, p = .009). It is also worth noting that creativity scores from both human judges and SemDis were consistently higher in the second (post-treatment) round than the first. This is in line with Mednick's (1962) associative theory and previous empirical findings that ideas tend to get more original and remote over time (e.g. Acar & Runco, 2014; Beaty & Silvia, 2012).

Overall, these results therefore indicated that if anything the lateral thinking treatment limited, rather than enhanced, people's fluency and creativity in this task. This was surprising, and closer qualitative analysis of participants' responses showed that they seldom seemed to utilise the random information with which they had engaged. In other words, it was relatively rare for participants who had read and summarised Wikipedia pages about certain sports or countries to subsequently provide ideas related to these subjects. This qualitative assessment suggested that participants sought to integrate a randomly presented subject into their response only if its relation to the question at hand was immediately apparent.

Study 3: Lateral thinking and divergent search II

Methods

85 participants (62.4% female) were randomly assigned to lateral thinking and control treatment conditions – identical to those used in Study 2 – and asked about three public policy proposals: state-subsidised abortion, the legalisation of cannabis, and the legalisation of medically assisted suicide (euthanasia). We selected these policies on the basis that there is substantial support both for and against them among the British public (ComRes, 2019).

The task mirrored the 'policy fluency task' used by Herzog and Hills (2019). For each policy, participants first stated their position on a seven-point scale (1 = in favor, 7 = against: see Supplemental Materials). They were then asked to imagine that they would be acting as an impartial mediator moderating a discussion on the policy. In this imaginary role, they were given two minutes to list all possible relevant arguments (in 10 words or less) that others might deem important for deciding in favor of or against the policy. Following this, participants received their respective treatment before being given a further two minutes to generate additional arguments for and against the same policy. Finally, they were asked to re-state their position on the same seven-point scale (without seeing their pre-treatment position).

We followed exactly the same method to score participants' responses for fluency and creativity as in Study 2, enabling us to similarly compare performances between rounds and conditions. We excluded responses from participants who were unable to generate more than one argument for a given policy. In addition, we excluded responses from those who failed to state their positions on a policy issue, or who clearly did not engage appropriately with the task. This left 69 participants for which we had valid responses for state-subsidised abortion, the legalisation of cannabis, and the legalisation of euthanasia.

Results

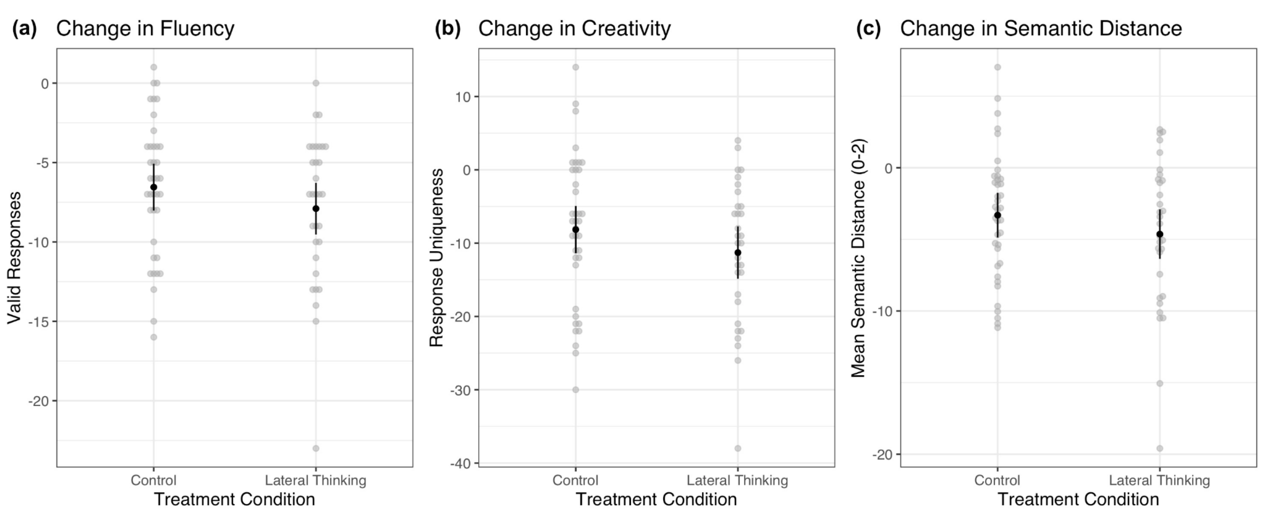
Our analysis of Study 3 data followed the same procedure as for Study 2, but with additional exploration of any treatment effect on changes in people's positions on each policy. We ran three ANCOVAs, controlling for pre-treatment scores in each: one to detect a treatment effect on post-treatment fluency; another to pick up an effect on post-treatment judge-scored creativity; and a third to detect an effect on post-treatment semantic distance creativity scores. This analysis, which included only participants who had generated valid responses for all three policy questions, highlighted a significant treatment effect on both fluency (p = .007) and creativity as scored by our judges (p = .036), but not on creativity as scored by the semantic distance model (p = 0.317). Once again contrary to our hypothesis, follow-up tests showed that participants in the control group generated an adjusted average of 7.8 valid post-treatment responses, compared with just 5.9 among those in the lateral thinking condition ($t(66) = 2.79$, $p < .01$). Similarly, the adjusted mean uniqueness (as scored by our human judges) of control group participants' post-treatment responses was also significantly higher than those in the lateral thinking condition ($t(66) = 2.14$, $p < .036$). The same trend was also evident in the creativity scores generated by SemDis on account of the strong correlation between these and the human judge scores ($r(67) = .76$, $p < .001$) although, as mentioned above, the difference between treatment groups was not significant. These results illustrated in Figure 3 (and summarised in Supplemental Table 4) indicated that our lateral thinking treatment proved more of a hindrance than a help to participants' fluency and creativity. In addition, it did not lead people to significantly shift their positions on public policy dilemmas in this task (see Supplemental Table 5).

Fig. 3. Change in Fluency and Creativity Scores, Study 3

Change in fluency (a), creativity scored by human judges (b), and creativity scored by the semantic distance model SemDis (c) between rounds. The plots show the distribution of participants' response fluency and creativity after the treatment relative to before. Grey points show the raw data, black points show the mean scores and error bars show 95% between-subjects confidence intervals.

Bayesian Meta-Analysis

To evaluate the overall main effect of the lateral thinking Wikipedia treatment, and to test support for the null hypothesis, we conducted a Bayesian meta-analysis (comparing the lateral thinking and control conditions only) using JASP software (JASP Team, 2020) to combine the results across the three studies (Wagenmakers et al, 2018). In addition, we ran Bayesian ANCOVAs using JASP (Rouder et al, 2012) to calculate Bayes Factors for each study. The models with the main effect of treatment on fluency and creativity provided overall support for the null hypothesis (Bayes factor for the null = BF_{01}, as opposed to support for the alternative = BF_{10}): BF_{10} = 0.24 (Study 1), BF_{10} = 0.82 (Study 2, fluency), BF_{10} = 0.53 (Study 2, judge-scored creativity), BF_{10} = 0.29 (Study 2, SemDis-scored creativity), BF_{10} = 0.47 (Study 3, fluency), BF_{10} = 0.51 (Study 3, judge-scored creativity), and BF_{10} = 0.43 (Study 3, SemDis-scored creativity). Put



differently, the data were 1/0.24 = 4.3 more likely to be observed under the null model in Study 1, etc.

For the meta-analysis, we used eta squared (η^2) as a standardised measure of effect size in each study because of its minimal absolute bias (Okada, 2013; results and equivalent plots using ω^2 can be found in Supplemental Materials). We calculated these values using the MOTE R package (Buchanan et al, 2019). The results of this analysis are shown in Figure 4 and summarised in Table 1, and provide compelling evidence in favor of the null hypothesis. Comparing the posterior probabilities for each model shows that the null hypothesis for the fixed effects model is almost 250x more likely than the alternative hypothesis. These results strongly suggest that our lateral thinking treatment had no meaningful effect on participants in the tasks involved in these studies.

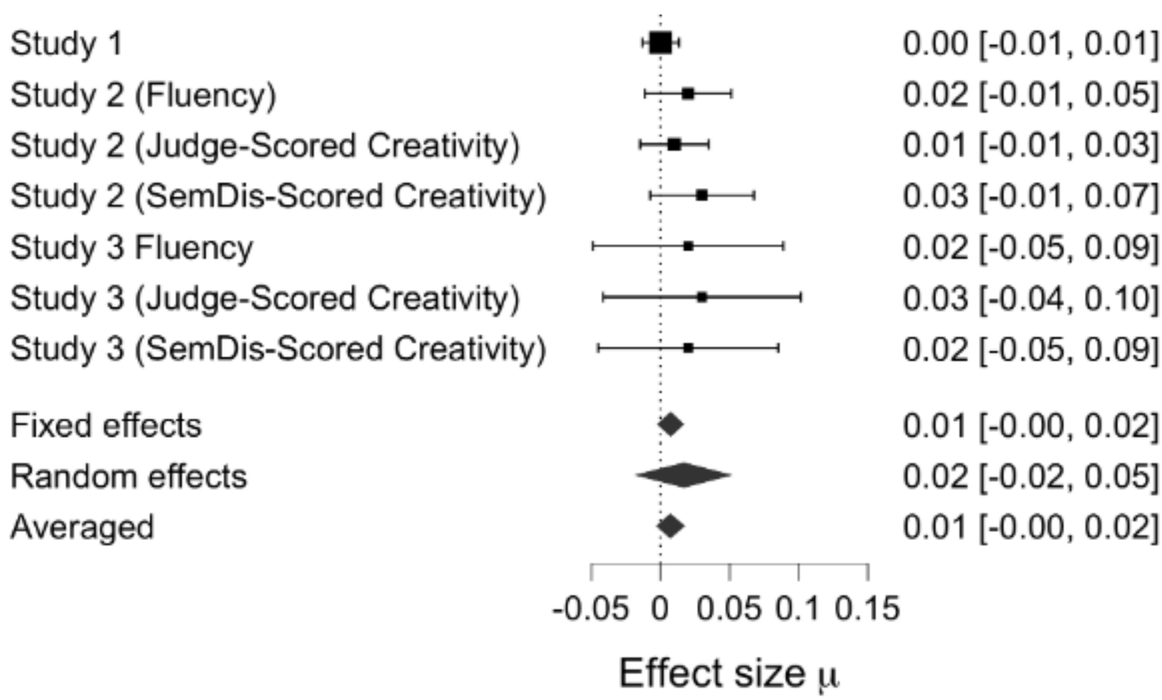


Fig. 4. Bayesian Meta-Analysis (All Studies)

Forest plot of the Bayesian meta-analysis showing the observed effect size (η^2) in each study and the estimated overall effect size per model. Positive effect sizes indicate the treatment is impaired relative to the control. Numbers in square brackets represent 95% lower and upper confidence bounds.

Table 1.
Prior and posterior model probabilities (ω^2)

	Prior	Posterior	BF
Fixed H_0	0.25	0.973	38.557
Fixed H_1	0.25	0.025	0.026

Note: Fixed H_0 refers to the fixed effects model null hypothesis, while H_1 represents the alternative hypothesis. BF refers to the Bayes Factor, which for the null hypothesis was considerably higher than for the random effects models.

Discussion

We anticipated that exposure to randomness would stimulate lateral thinking and in turn improve participant's performance in both convergent and divergent tasks. This hypothesis was based on previous suggestions and research indicating that randomness might help people to generate more ideas and avoid local maxima when sampling from memory (e.g., Beaney, 2005; Campbell, 1960; De Bono, 2010; Vul & Pashler, 2008; Christian & Griffiths, 2016). In turn, we theorised that these benefits would support participants to make more accurate forecasts and generate more creative responses. However, our lateral thinking treatment of random Wikipedia pages demonstrated no consistent positive effect in the convergent task (Study 1) and seemed more of a hindrance than a help in the divergent tasks (Studies 2 and 3). A Bayesian meta-analysis found strong support for the null hypothesis.

Previous research has found that when minds cannot find what they are looking for, they engage in relatively more exploratory behavior, similar to the way animals engage in increasingly more exploration when they are unable to find what they are looking for (Hills, 2019). The hungry cats in Thorndike's (1898) puzzle boxes provide one of the earliest and most vivid illustrations of this tendency. There is also evidence that minds engage in more high-temperature exploratory search in their internal environment (i.e. memory) to solve both convergent and divergent search problems (Hills et al, 2015). This modulation of exploration and exploitation is consistent with a great deal of research indicating its adaptive value in response to changing environments (Yang, 2020; Gopnik, 2020; Hills, Kalff, & Wiener, 2013; Zhang, Sjoerds, & Hommel, 2020).

Why then did we not observe an influence of random stimuli and how does this result inform us about their utility? Revisiting the potential processes by which our lateral thinking treatment might work, we suggest two possibilities: One possibility is that random stimuli helps induce a style of thinking that is more explorative--changing the search dynamic inside peoples' heads. Bombarded with random stimuli, cognition might engage in more high temperature, exploratory thinking, relaxing its tendencies to exploit more well worn patterns of thought. This does not appear to be supported by our data.

A second possibility is that random stimuli may, by chance, prime participants to think about specific topics that turn out to be relevant even though they were not previously anticipated. In this case, it is not the cognitive process that is induced into greater exploration, but an external process that generates exploratory stimuli that sometimes either turn out to be relevant or to point the way towards more relevant ideas. This suggests the need for a degree of relative randomness (i.e., temperature modulation) that is more appropriate to distribution of potentially relevant stimuli, similar to optimally random search algorithms (Wei, Chen, Yu, & Chen, 2019). Our Wikipedia treatment may have hindered participants' task performance because it introduced a degree of randomness that was too high, and remained so throughout the task; the treatment may have been unhelpful because it was *too random,* rarely generating Wikipedia pages with clear connections to the task at hand. In practice, random pages relating to public health, epidemics, or even COVID-19 itself may have proved more helpful for participants in this experiment (Wikipedia does provide a random-in-category search: https://en.wikipedia.org/wiki/Special:RandomInCategory). To add insult to injury, unsuccessful efforts to find connections may have derailed promising trains of thought – a disruption that seemed not to affect participants completing mathematical additions in the control condition, who are unlikely to have even attempted to seek connections between the two tasks.

To serve as a decision aid, random stimuli may need to fall within a Goldilocks zone of constrained relative randomness. This follows a similar argument for constrained stochastic behavior in relation to creativity (Simonton, 2004). Not enough randomness means our thinking might remain firmly 'inside the box'. Too much randomness may help us escape the box, but it might also lead to thinking unrelated to the original problem. This may help to explain the potential power of some ritual problem solving methods such as the I Ching and Tarot: by offering guidance that is both abstract and relatable, they might more frequently occupy the zone of optimal randomness and provide interpretations of problems that are at least plausible (Jung, 1950).

Interpreting our results through the lens of optimal randomness, inducing lateral thinking may require random stimuli that are nonetheless related to potential solutions--even though this may be difficult to predict--and that unrelated or "irrelevant" stimuli are unlikely to help people think laterally or unravel problems in either a divergent or convergent way. Of course, one can argue that since it is not clear how to predict the relevance of random stimuli in advance, the problem may not be with random stimuli per se, but with the amount of random stimuli needed to find the Goldilocks stimuli--those that are both unpredictable and yet still relevant. In that sense, prior claims about the value of random stimuli may still be true in the sense that sometimes one gets lucky; sometimes apparently irrelevant stimuli are in fact relevant, but one does not know until they look. Our results suggest this might be rarer than previously expected and that, on the contrary, if one wants to think outside the box, it might make sense to at least try to think nearby.

Acknowledgements

This research was funded by the University of Warwick.

References

Acar, S., & Runco, M. A. (2014). Assessing associative distance among ideas elicited by tests of divergent thinking. *Creativity Research Journal*, *26*(2), 229-238.

Acar, S., & Runco, M. A. (2015). Thinking in multiple directions: Hyperspace categories in divergent thinking. *Psychology of Aesthetics, Creativity, and the Arts*, *9*(1), 41.

Ansburg, P. I., & Dominowski, R. I. (2000). Promoting insightful problem solving. *The Journal of Creative Behavior*, *34*(1), 30-60.

Beaney, M. (2005). *Imagination and creativity* (Vol. 4). Open University Worldwide Ltd.

Beaty, R. E., & Silvia, P. J. (2012). Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(4), 309.

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior research methods*, *53*(2), 757-780.

Buchanan, E. M., Gillenwaters, A., Scofield, J. E., & Valentine, K. (2019). MOTE: Measure of the Effect: Package to assist in effect size calculations and their confidence intervals. *R package version*, *1*(2).

Byrne, C. L., Shipman, A. S., & Mumford, M. D. (2010). The effects of forecasting on creative problem-solving: An experimental study. *Creativity Research Journal*, *22*(2), 119-138.

Campbell, D.T. (1960). Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological review*, *67*(6), 380.

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2003). Decreased latent inhibition Is associated with increased creative achievement in high-functioning individuals. Journal of Personality and Social Psychology, *85(3)*, 499–506.

Christian, B. and Griffiths, T. (2016). *Algorithms to live by: The computer science of human decisions*. Macmillan.

Chrysikou, E. G. (2019). Creativity in and out of (cognitive) control. *Current Opinion in Behavioral Sciences*, *27*, 94-99.

ComRes. (2019). *BBC Beliefs Survey - November 2019*. Retrieved from https://comresglobal.com/polls/bbc-beliefs-survey-november-2019/

Cropley, A. (2006). In praise of convergent thinking. *Creativity research journal*, *18*(3), 391-404.

De Bono, E. (2010). *Lateral thinking: a textbook of creativity*. Penguin UK.

de Vries, H. B., & Lubart, T. I. (2019). Scientific creativity: divergent and convergent thinking and the impact of culture. *The Journal of Creative Behavior*, *53*(2), 145-155.

Dumas, D., Organisciak, P., & Doherty, M. (2020). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). Creative cognition: Theory, research, and applications.

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B*, *375*(1803), 20190502.

Guilford, J. P. (1950). Fundamental statistics in psychology and education.

Guilford, J. P. (1967). The nature of human intelligence.

Hanke, J. E., Reitsch, A. G., & Wichern, D. W. (2001). *Business forecasting* (Vol. 9). New Jersey: Prentice Hall.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231-237.

Herzog, S. M., & Hills, T. T. (2019). Mediation Centrality in Adversarial Policy Networks. *Complexity*, *2019*.

Hills, T. T. (2019). Neurocognitive free will. *Proceedings of the Royal Society B*, *286*(1908), 20190510.

Hills, T. T., Kalff, C., & Wiener, J. M. (2013). Adaptive Lévy processes and area-restricted search in human foraging. *PLoS One*, *8*(4), e60488.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., & Cognitive Search Research Group. (2015). Exploration versus exploitation in space, mind, and society. *Trends in cognitive sciences*, *19*(1), 46-54.

Hommel, B. (2012). Convergent and divergent operations in cognitive search. *Cognitive search: Evolution, algorithms, and the brain*, 221-235.

Hubbard, D. W. (2020). *The failure of risk management: Why it's broken and how to fix it*. John Wiley & Sons.

JASP Team (2020). JASP (Version 0.14) [Computer software].

Jung, C. (1950). “Foreword to the ‘I Ching.’” In *Psychology and Religion: West and East, Collected Works 11.* Princeton, N.J.: Princeton University Press.

Katsagounos, I., Thomakos, D. D., Litsiou, K., & Nikolopoulos, K. (2020). Superforecasting reality check: Evidence from a small pool of experts and expedited identification. *European Journal of Operational Research*.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *science*, *220*(4598), 671-680.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of personality and social psychology*, *47*(6), 1231.

Maio, S., Dumas, D., Organisciak, P., & Runco, M. (2020). Is the reliability of objective originality scores confounded by elaboration?. *Creativity Research Journal*, *32*(3), 201-205.

[Dataset] Malthouse, E., Hills, T. T., (2021), *Exploratory Divination in Decision Making / Data and Code*, accessible at: https://osf.io/pzgjx/

Mednick, S. (1962). The associative basis of the creative process. *Psychological review*, *69*(3), 220.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much?. *The quarterly journal of economics*, *122*(3), 1067-1101.

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika*, *40*(2), 129-147.

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22-27.

Plucker, J. A., & Makel, M. C. (2010). Assessment of creativity.

Proctor, R. A. (1993). Computer stimulated associations. *Creativity Research Journal*, *6*(4), 391-400.

Runco, M. A. (2010). Divergent thinking, creativity, and ideation. *The Cambridge handbook of creativity*, *413*, 446.

Russo, J. E., & Schoemaker, P. J. (1992). Managing overconfidence. *Sloan management review*, *33*(2), 7-17.

Simonton, D. K. (2011). Creativity and discovery as blind variation: Campbell's (1960) BVSR model after the half-century mark. *Review of General Psychology*, *15*(2), 158-174.

Simonton, D. K. (2004). Creativity as a Constrained Stochastic Process. In R. J. Sternberg, E. L. Grigorenko, & J. L. Singer (Eds.), *Creativity: From potential to realization* (pp. 83–101). American Psychological Association.

Sloane, P. (2017). *The Leader's Guide to Lateral Thinking Skills: Unlock the Creativity and Innovation in You and Your Team*. Kogan Page Publishers.

Sweller, J. (2009). Cognitive bases of human creativity. *Educational Psychology Review*, *21*(1), 11-19.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Thorndike, E. L. (1898). Animal intelligence: an experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i.

Vul, E. and Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645-647.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... & Meerhoff, F. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic bulletin & review*, *25*(1), 58-76.

Wei, J., Chen, Y., Yu, Y., & Chen, Y. (2019). Optimal Randomness in Swarm-Based Search. *Mathematics*, *7*(9), 828.

Weisberg, R. W., & Markman, A. (2009). On “out-of-the-box” thinking in creativity. *Tools for innovation*, 23-47.

Wertheimer, M., & Wertheimer, M. (1959). *Productive thinking*. New York: Harper.

Wilson, R. C., Guilford, J. P., & Christensen, P. R. (1953). The measurement of individual differences in originality. *Psychological bulletin*, *50*(5), 362.

Yang, X. S. (2020). *Nature-inspired optimization algorithms*. Academic Press.

Zhang, W., Sjoerds, Z., & Hommel, B. (2020). Metacontrol of human creativity: The neurocognitive mechanisms of convergent and divergent thinking. *NeuroImage*, *210*, 116572.

Supplemental Materials

Data

All data and analysis scripts in R can be accessed at: https://osf.io/pzgjx/

Survey Questions

Study 1

Forecast 1a: *The number of deaths in the world attributed to COVID-19 currently stands at 649,358, as of 26th July 2020. What do you predict this figure will be on 10th August 2020 (two weeks from now)? Please provide a best estimate, a lower and an upper bound, such that you are 90% confident that the true value will lie in between these lower and upper bound figures.*

Forecast 1b: *What do you predict this figure will be by 31st July 2021 (one year from now)? As above, please provide a best estimate, a lower and an upper bound, such that you are 90% confident that the true value will lie in between these lower and upper bound figures.*

Forecast 2a: *As of 26th July 2020, 22 countries had reported more than 5,000 COVID-related deaths. How many countries do you predict will have reached this unfortunate milestone by 10th August 2020 (two weeks from now)? Please provide a best estimate, a lower and an upper bound, such that you are 90% confident that the true value will lie in between these lower and upper bound figures.*

Forecast 2b: *How many countries do you predict will have reached this unfortunate milestone by 31st July 2021 (one year from now)? As above, please provide a best estimate, a lower and an upper bound, such that you are 90% confident that the true value will lie in between these lower and upper bound figures.*

Forecast 3a: *Unemployment measures people without a job who have been actively seeking work within the last four weeks and are available to start work within the next two weeks. Between January and March 2020, the UK unemployment rate was estimated to be 3.9%. What do you predict this figure will be for the period covering March – May 2020 (this year)? As before, please provide a best estimate, a lower and an upper bound, such that you are 90% confident that the true value will lie in between these lower and upper bound figures.*

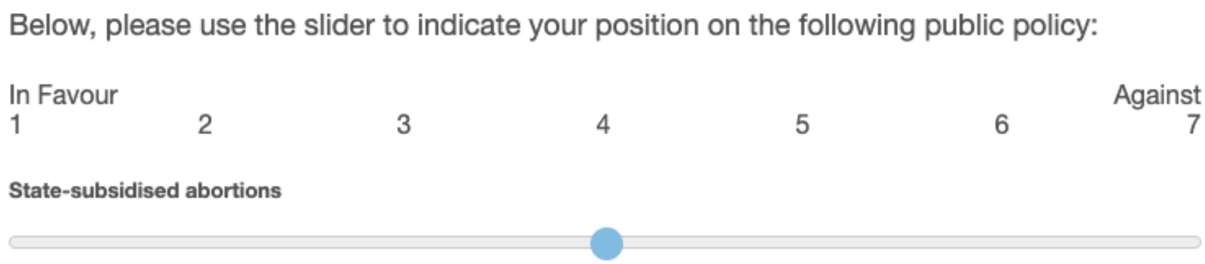
Forecast 3b: *What do you predict this figure will be for the period covering March – May 2021 (next year)? As above, please provide a best estimate, a lower and an upper bound, such that you are 90% confident that the true value will lie in between these lower and upper bound figures.*

Study 2

Round 1: *Please list as many predictions as you can about how society might change after the COVID-19 pandemic.*

Round 2: *Taking the subjects covered in the previous page into account, please generate as many additional predictions as you can about how society might change after the COVID-19 pandemic.*

<u>Study 3</u>



Please imagine that you will be acting as an impartial mediator moderating a discussion on the above policy.

You will be asked to list all the arguments for and against the above policy that you can think of that other people might find important for deciding in favor of or against the policy.

Treatment Examples

<u>Lateral Thinking</u>

Coal power in the United States

From Wikipedia, the free encyclopedia

See also: Coal mining in the United States

Coal power in the United States accounted for 39% of the country's electricity production at utility-scale facilities in 2014, 33% in 2015, 30.4% in 2016, 30.0% in 2017, 27.4% in 2018, and 23.5% in 2019.[2] Coal supplied 12.6 quadrillion Btu (3,700 TWh) of primary energy to electric power plants in 2017, which made up 91% of coal's contribution to US energy supply.[3] Utilities buy more than 90% of the coal consumed in the United States.[4]

Coal has been used to generate electricity in the United States since an Edison plant was built in New York City in 1882.[5] The first AC power station was opened by General Electric in Ehrenfeld, Pennsylvania in 1902, servicing the Webster Coal and Coke Company.[5] By the mid-20th century, coal had become the leading fuel for generating electricity in the US. The long, steady rise of coal-fired generation of electricity shifted to a decline after 2007. The decline has been linked to the increased availability of natural gas, decreased consumption,[6] renewable power, and more stringent environmental regulations. The Environmental Protection Agency has advanced restrictions on coal plants to counteract mercury pollution, smog, and global warming.

Contents [hide]

Sources of Electricity Generation United States - 2016

Sources of electricity for 2016.[1]

Coal electrical generation (black line), compared to other sources, 1949-2016

Please briefly summarise the topic of the above page in 10 words or less.

Consider-the-opposite

Think back to your estimates about the future number of worldwide deaths attributed to COVID-19.

First, assume that your previous estimates are off the mark. Second, think about a few reasons why that could be. Which assumptions and considerations could have been wrong? Third, think about what these new considerations mean – were your first estimates too low or too high?

- My first estimates were too low
- My first estimates were too high
- Not sure

Control

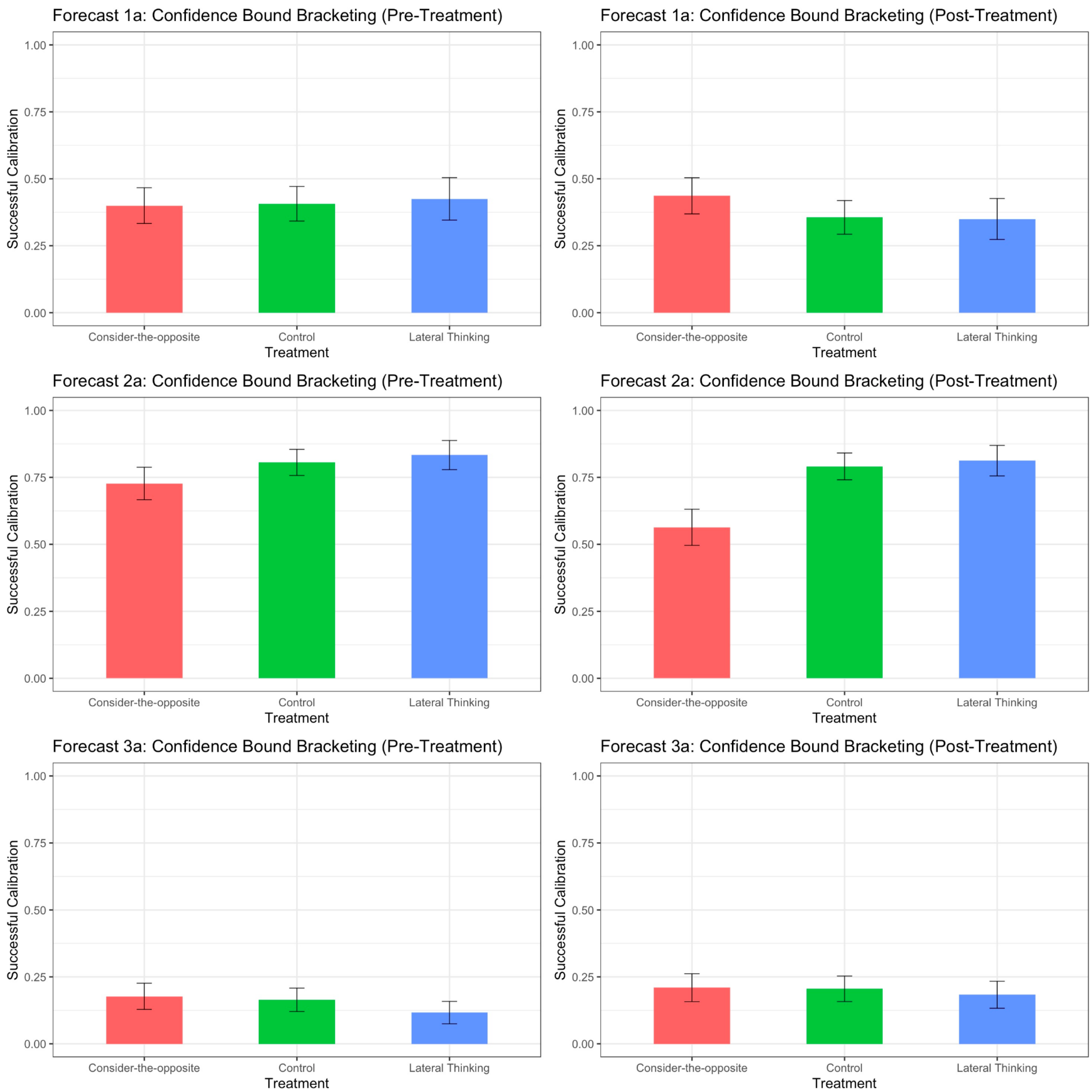
20 + 72 + 53 + 60 + 24 = ?

91 + 35 + 39 + 12 + 93 = ?

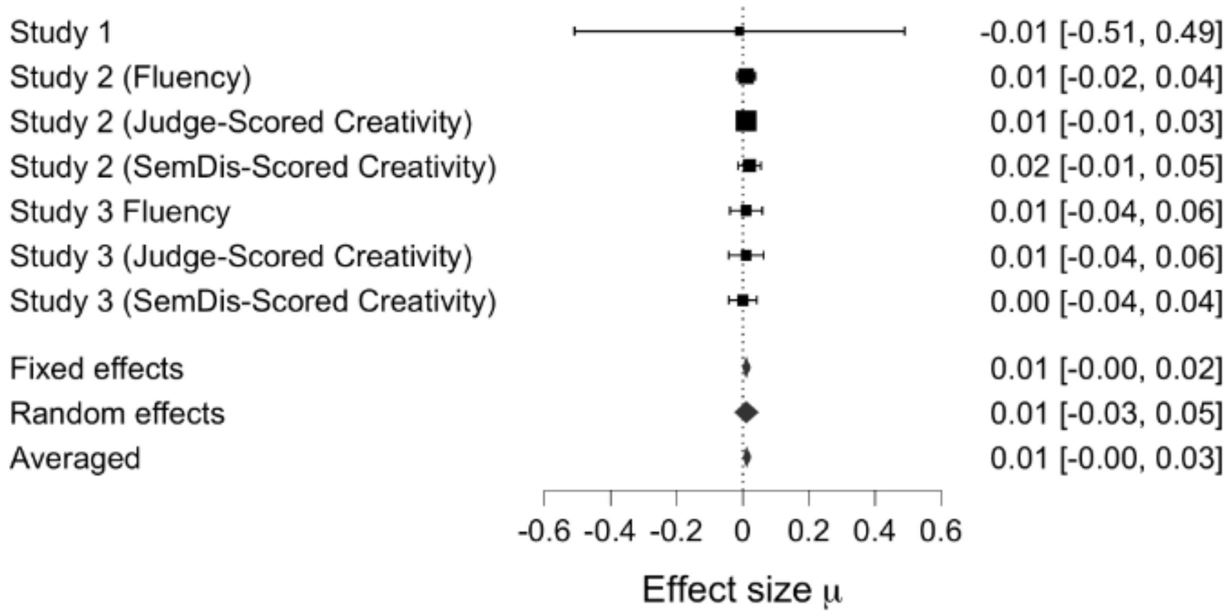
61 + 93 + 19 + 43 + 42 = ?

35 + 30 + 64 + 61 + 52 = ?

Supplemental Figures



Supplemental Fig. 1.
Confidence bound calibration, defined as the proportion of 90% bound estimates that successfully bracketed the true value, before and after treatments. Responses are divided by forecast and condition, with standard error bars shown. Forecast 3a graphs illustrate the slight improvement in estimates for this question, unlike for Forecasts 1a and 2a.



Supplemental Fig. 2.

Forest plot from Bayesian meta-analysis, with effect size as ω^2. Numbers in square brackets represent 95% lower and upper confidence bounds.

1 **<u>Supplemental Tables</u>**

Table 1.
Summary of interactions between treatment and demographic variables and pre-treatment scores. No significant interactions were found, justifying a subsequent ANCOVA with Type II sums of squares.

Study	Dependent Variable	Treatment	Age	Gender
Study 1	Forecast Accuracy	$F(2, 126) = 0.3$, $p = .74$	$F(4, 124) = 0.84$, $p = .50$	$F(2, 126) = 0.33$, $p = .72$
Study 2	Fluency	$F(1, 265) = 0.18$, $p = .67$	$F(5, 261) = 0.43$, $p = .82$	$F(4, 262) = 1.23$, $p = .30$
	Creativity	$F(1, 265) = 0$, $p = .96$	$F(5, 261) = 0.27$, $p = .93$	$F(4, 262) = 0.73$, $p = .57$
Study 3	Fluency	$F(1, 67) = 0.50$, $p = .48$	$F(4, 64) = 1.33$, $p = .27$	$F(1, 67) = 0.06$, $p = .80$
	Creativity	$F(1, 67) = 0.48$, $p = .49$	$F(4, 64) = 2.05$, $p = .10$	$F(1, 67) = 0.01$, $p = .93$

2

Table 2.
Study 1: first (pre-treatment) and revised (post-treatment) forecast errors (defined as mean absolute percentage error: MAPE) by forecast and treatment condition.

Forecast	Main Effect: Treatment	First Forecast Error (MAPE)			Revised Forecast Error (MAPE)		
		Consider the opposite	Lateral Thinking	Control	Consider the opposite	Lateral Thinking	Control
1	$p = .057$	6.3	6.2	11.4	8.1	7.0	8.1
2	$p = .01$	25.1	12.3	22.1	31.9	15.7	14.9
3	$p = .508$	98.3	123.8	90.2	90.4	106.0	66.4

3

Table 3.
Study 2: Mean Semantic Distance Scores for participant responses aggregated by round and treatment condition. The score for each individual response is computed as the mean of semantic distance scores across 5 semantic spaces (GloVe, TASA, cbowbaroni, cbowsubtitle, cbowukwacsubtitle). Scores range from 0 to 2 in all semantic models, where higher scores indicate greater semantic distance between the subject (pandemic) and a response.

Treatment	Mean Semantic Distance Score (0-2)		Difference	*p*-value
	Round 1	Round 2		
Lateral Thinking	0.984	0.987	0.003	.147
Control	0.985	0.988	0.003	.131

Table 4.
Study 3: summary of mean number of valid responses before and after the treatment in each condition. Individual responses were rated blindly as either valid (1) or not valid (0) by three independent judges.

Treatment	Abortion: valid responses (before)	Abortion: valid responses (after)	Cannabis: valid responses (before)	Cannabis: valid responses (after)	Euthanasia: valid responses (before)	Euthanasia: valid responses (after)
Lateral Thinking	4.12	1.68	4.94	2.35	4.39	1.73
Control	4.31	2.43	4.91	2.93	4.74	2.32

Table 5.
Mean positions on each policy before and after the treatments and engaging in the process of generating arguments for and against the policy in question. Responses fall on a scale of 1-7 (1 = in favor, 7 = against). In each instance, participants tended to become slightly more in favor of the policies, but differences in movements between treatment groups were not significant.

Treatment	Abortion			Cannabis			Euthanasia		
	Initial Position	Revised Position	Change (+/-)	Initial Position	Revised Position	Change (+/-)	Initial Position	Revised Position	Change (+/-)
Lateral Thinking	2.44	2.21	-0.24	3.24	2.94	-0.29	2.76	2.70	-0.06
Control	2.86	2.40	-0.45	3.51	3.44	-0.07	2.87	2.74	-0.13

Table 6.
Prior and posterior model probabilities with effect size measured as η^2

	Prior	Posterior	BF_{10}
Fixed H_0	0.25	0.987	162.347
Fixed H_1	0.25	0.006	0.006

Note: Fixed H_0 refers to the fixed effects model null hypothesis, while H_1 represents the alternative hypothesis. BF_{10}

refers to the Bayes Factor for both models, which for the null hypothesis was considerably higher than for the random effects models.